Tijdschrift voor Medisch Onderwijs, januari 2001 | Vol. 20, nr.1, p. 17-24

Bepaling van de referentiewaarde bij screenende kennistoetsen door het NHG met de methode van Angoff

J. ter Burg, W.H.M. Emons, S.S.L. Mol, L.W.T. Schuwirth

Samenvatting

Inleiding: *Om huisartsen feedback te kunnen geven over de mate waarin zij de richtlijnen uit een standaard van het Nederlands Huisartsen Genootschap (NHG) beheersen, moet een referentiewaarde beschikbaar zijn. Onderzocht is of de methode van Angoff een geschikte methode is om de referentiewaarde te bepalen.*

Methode: *Met behulp van een Angoff procedure is de referentiewaarde door zes experts bepaald voor drie subtoetsen over Astma en COPD.*

Resultaten: *De betrouwbaarheid van de beoordelingen is hoog en bereikt 0.80 als acht experts de items beoordelen. De geloofwaardigheid van de absolute referentiewaarden is onderzocht door de correlatie tussen de p-waarde geschat door de experts en de p-waarde geobserveerd bij toetsafname te bepalen. Deze bleek voldoende te zijn (0.78-0.85). De praktische uitvoerbaarheid, onder andere in termen van tijdbelasting en helderheid van instructies, bleek goed.*

Conclusie: *Op basis van de resultaten lijkt de conclusie gerechtvaardigd dat de methode van Angoff geschikt is voor bepaling van de referentiewaarde bij een screenende toets over een NHG-standaard. (Burg J ter, Emons WHM, Mol SSL, Schuwirth LWT. Bepaling van de referentiewaarde bij screenende kennistoetsen door het NHG met de methode van Angoff. Tijdschrift voor Medisch Onderwijs 2001;20(1):17-24.)*

Inleiding

Bij vrije keuze van nascholing, bestaat de kans dat huisartsen onderwerpen kiezen waarmee zij affiniteit voelen en waarin zij al vrij goed zijn.[1] Een instrument waarmee men objectief kan vaststellen op welke gebieden men minder goed is, kan mogelijk stimuleren om zich juist op die gebieden te gaan nascholen. Het NHG ontwikkelt voor huisartsen een dergelijk instrument: de zogenaamde screenende toets. Op dit moment zijn alleen screenende toetsen beschikbaar voor kennis over NHG-standaarden.

Om huisartsen feedback te kunnen geven over de mate waarin zij de richtlijnen uit een standaard beheersen, moet er een grensscore beschikbaar zijn. De toetsscore die de grens tussen voldoende en onvoldoende kennis aangeeft, zal in het vervolg van dit artikel met referentiewaarde aangeduid worden. In de onderwijskundige literatuur wordt vaak gesproken over de cesuur.

Een referentiewaarde kan op drie verschillende manieren bepaald worden: vanuit het leerdomein (absolute methode), vanuit de groepsresultaten (relatieve methode) of vanuit een extern criterium.[2] In het verleden heeft het NHG gebruik gemaakt van een relatieve methode: als norm werd gehanteerd het gemiddelde van de toetsscores van alle deelnemende huisartsen minus tweemaal de standaard-meetfout van het gemiddelde. Groepen huisartsen met een score onder deze norm kregen het advies nascholing te volgen. Ondanks de plausibiliteit van deze methode zijn er ook bezwaren aan te voeren. Bij de relatieve methode kan de referentiewaarde alleen achteraf vastgesteld worden, omdat er empirische gegevens nodig zijn. De organi-

satie voor het werven van deelnemers, het verzenden van de toetsen, het uitvoeren van de analyses en het genereren van de feedback blijkt zeer arbeidsintensief te zijn. Ook is het de vraag of de gemiddelde score wel overeenkomt met het gewenste kennisniveau. In het kader van kwaliteitsbewaking kan verdedigd worden dat voor een huisarts een bepaald niveau van kennis belangrijker is dan evenveel kennis als, of meer kennis dan, de gemiddelde huisarts. Deze bezwaren pleiten ervoor een absolute referentiewaarde te gebruiken, een referentiewaarde die voorafgaand aan de toetsafname vastgesteld kan worden en onafhankelijk van de toetsresultaten is.

Voor het bepalen van een absolute referentiewaarde is een groot aantal procedures ontwikkeld. Op basis van de literatuur en ervaringen van experts is in dit onderzoek gekozen voor de methode van Angoff. Uit onderzoek van Berk komt naar voren dat deze methode de beste balans vormt tussen enerzijds de praktische haalbaarheid en anderzijds de technische adequaatheid. Daarnaast is de methode van Angoff volgens Berk conceptueel eenvoudig voor te stellen en breed toepasbaar.[3] Het is dan ook niet verwonderlijk dat de methode van Angoff een veelgebruikte methode binnen onderwijsevaluatie en certificering is.[4 5] Ook binnen het medisch onderwijs wordt de Angoff-methode veelvuldig toegepast.[6-8]

Dit onderzoek is gericht op het vaststellen van de betrouwbaarheid en de overeenstemming van de expertbeoordelingen, de geloofwaardigheid en de praktische uitvoerbaarheid van de procedure en het oordeel van de experts over de methode van Angoff. Op basis van de resultaten zal gekeken worden in hoeverre de methode van Angoff geschikt is en zullen eventuele verbeteringen worden voorgesteld.

Methode

Toets

De NHG-kennistoets 'COPD/Astma' bestaat uit 56 items verdeeld over drie subtoetsen waarin respectievelijk de NHG-standaarden 'COPD/Astma: diagnostiek', 'Astma: behandeling' en 'COPD: behandeling' getoetst worden. Voor elke subtoets is de referentiewaarde bepaald.

Elke subtoets bestaat uit een aantal casus met één of meer stellingen (figuur 1). Een huisarts moet per stelling aangeven of deze juist of onjuist is. Als een huisarts het antwoord niet weet of twijfelt, kan de vraagtekenoptie gebruikt worden. De toetsscores worden berekend volgens de zogenaamde 'percentage goed-min-fout'-scoreberekening. Het aantal foute antwoorden wordt van het aantal goede antwoorden afgetrokken, het verschil wordt gedeeld door het totale aantal items en vermenigvuldigd met honderd.

Methode van Angoff

Bij de methode van Angoff wordt een aantal experts gevraagd om bij elk item de kans te schatten dat een minimaal compe-

Figuur 1. *Voorbeeld van een casus met stelling.*

De heer Jonkers, 48 jaar, komt op het spreekuur van de huisarts omdat hij hoest. Hij is niet bekend met astma en rookt circa dertig sigaretten per dag. Na anamnese en lichamelijk onderzoek twijfelt de huisarts tussen astma en COPD. Hij besluit tot het doen van aanvullend onderzoek.

1. In het geval van de heer Jonkers heeft piekstroommeting de voorkeur boven spirometrie.

❍ juist ❍ onjuist ❍ ?

tente kandidaat – een kandidaat die net op de grens tussen voldoende en onvoldoende zit – het item correct beantwoordt. Door de succeskansen bij elkaar op te tellen, wordt de verwachte somscore voor de minimaal competente kandidaat op de toets berekend. De uiteindelijke referentiewaarde is het gemiddelde over de expertbeoordelingen.[9]

In dit onderzoek is de minimaal competente kandidaat gedefinieerd als een *normaal goed functionerende huisarts*. Onder een normaal goed functionerende huisarts wordt een huisarts verstaan die door nascholing zijn of haar kennis, vaardigheden en praktijkvoering in voldoende mate op peil houdt. Getracht is de beoordelingstaak eenvoudiger te maken door de experts te vragen te schatten hoeveel huisartsen van een honderdtal normaal goed functionerende huisartsen het item correct zullen beantwoorden, in plaats van ze te vragen de succeskans van een enkel individu op een item te schatten.

De methode van Angoff is ontwikkeld voor het bepalen van de referentiewaarde voor een toets met dichotome items, dat wil zeggen items die goed of fout gescoord worden. In dit onderzoek willen we echter de referentiewaarde bepalen voor polytome items. Besloten is om dicht bij de methode van Angoff te blijven en de items als dichotoom te beschouwen. De experts werd alleen gevraagd het percentage dat het item correct beantwoordt te schatten. Het afgeronde gemiddelde over de expertbeoordelingen levert het minimale aantal items dat goed beantwoord moet worden; de overige items mogen fout beantwoord worden. Bij het bepalen van de referentiewaarde wordt de mogelijkheid dat een respondent de vraagtekenoptie invult dus genegeerd. De referentiewaarde uitgedrukt als 'percentage goed-min-fout' kan nu eenvoudig berekend worden.

Het expertpanel

Het expertpanel bestond uit zes huisartsen: twee huisartsen uit de werkgroep die de NHG-standaard opgesteld heeft, twee huisarts-stafleden van de NHG-afdeling Standaarden en twee huisarts-stafleden van de NHG-afdeling Deskundigheidsbevordering. De eerste vier experts zijn nauw betrokken geweest bij de ontwikkeling van de betreffende NHG-standaarden, de laatste twee bij het ontwerp en de ontwikkeling van de nascholingsmaterialen over de betreffende onderwerpen. Bovendien zijn de meeste experts werkzaam in een huisartsenpraktijk.

Evaluatie

De verkregen referentiewaarden zijn geëvalueerd aan de hand van de toetsscores van 289 huisartsen die de toets in 1998 maakten. Van de 289 huisartsen hebben 257 huisartsen de toets onder examencondities gemaakt, dat wil zeggen dat zij tijdens de beantwoording van de items naslagwerken noch elkaar konden raadplegen. De resterende 32 huisartsen – de 'voorbereiders' – hebben zich voorbereid door één of meer NHG-standaarden, een leerboek of ander naslagwerk voorafgaand aan de toetsafname door te nemen of tijdens de toetsafname te raadplegen.

Moeilijkheidsgraad versus relevantie

Uit de pretest met een toets over een ander onderwerp bleek dat experts in veel gevallen een relevantieoordeel gaven, terwijl hun gevraagd was een schatting van de moeilijkheidsgraad te geven. Om verwarring tussen relevantie en moeilijkheid van een item te voorkomen, is de volgende constructie bedacht: een expert wordt eerst gevraagd een oordeel over de relevantie te geven; daarna wordt een schatting van de moeilijkheidsgraad gevraagd. Opgemerkt dient te worden dat de relevantieoordelen *niet* meegenomen zijn bij de bepaling van de referentiewaarde.

Na afloop van de pretest is de experts gevraagd hun mening over de methode te geven. Enkele experts gaven aan dat zij een zekere weerstand moesten overwinnen om de moeilijkheid te schatten van een item waarvan zij de kwaliteit niet hoog achtten. Om deze reden is besloten de experts in de gelegenheid te stellen inhoudelijke op- of aanmerkingen bij een casus of de bijbehorende items te maken.

Twee rondes

De procedure werd uitgevoerd in twee rondes. In de eerste ronde kreeg elke expert een antwoordformulier. Op het antwoordformulier staan de casus en de bijbehorende items weergegeven. De juiste antwoorden en de toelichting op de antwoorden zijn in een aparte bijlage gegeven. Bij elk item werd de expert gevraagd de relevantie te beoordelen, de moeilijkheidsgraad te schatten en eventueel commentaar op het item te geven. Daarnaast werd de experts gevraagd een toelichting op het geschatte succespercentage te geven. In de tweede ronde, die als doel had de mate van overeenstemming tussen de experts te vergroten, werd de experts gevraagd de items waarvan de geschatte itemmoeilijkheden relatief sterk uiteenliepen (sd. ≥ 0.18) opnieuw te beoordelen. De experts ontvingen naast hun eigen beoordelingen en toelichtingen uit de eerste ronde ook de beoordelingen en de toelichtingen van de andere experts. Beide rondes waren schriftelijk.

Analyses

De betrouwbaarheid van de expertbeoordelingen werd bepaald met Cronbachs alfa. Cronbachs alfa geeft de mate aan waarin de geschatte succespercentages correleren. De interbeoordelaarsovereenstemming werd onderzocht met de phi-coëfficiënt. Als de beoordelingen van verschillende beoordelaars perfect overeenstemmen, dat wil zeggen per item gelijk zijn, dan is de coëfficiënt gelijk aan 1.[10] Voor zowel Cronbachs alfa als de phi-coëfficiënt geldt dat een waarde van rond de 0.80 in het algemeen als voldoende beschouwd wordt.[2 10]

Om een beeld te krijgen van het aantal beoordelaars dat nodig is om de gewenste overeenstemming van 0.80 te verkrijgen, is de Spearman-Brown-formule gebruikt. Om na te gaan of de methode geloofwaardig is, werd ten eerste de samenhang tussen de geobserveerde itemmoeilijkheden (uit de afname bij 289 huisartsen) en de geschatte itemmoeilijkheden (uit de expertprocedure) bepaald. Een sterke samenhang draagt immers bij aan de geloofwaardigheid van de gebruikte procedure, omdat deze erop wijst dat de experts blijkbaar in staat zijn de moeilijke en de makkelijke items van elkaar te onderscheiden. Ten tweede werd de groep 'voorbereiders' vergeleken met de groep 'niet-voorbereiders' wat betreft het aantal respondenten dat onder de referentiewaarde scoort en het aantal respondenten dat boven de referentiewaarde scoort. Hierbij werd uitgegaan van de veronderstelling dat een huisarts die voorafgaand aan de toetsafname de betreffende NHG-standaard doorgenomen heeft, hoger scoort dan de vastgestelde referentiewaarde. Als een hoog percentage 'voorbereiders' niet boven de vastgestelde referentiewaarde scoort, lijkt de referentiewaarde niet reëel. Ten derde werd gekeken of er sprake was van een zeer hoge dan wel een zeer lage referentiewaarde, door het percentage respondenten dat onder, respectievelijk boven de referentiewaarde scoort te vergelijken met het percentage respondenten dat onder respectievelijk boven de gemiddelde toetsscore van alle respondenten scoort. Om te bepalen of de methode praktisch uitvoerbaar is, is na afloop van elke ronde een korte vragenlijst afgenomen over de benodigde tijd voor de beoordelingen, duidelijkheid van de instructies, ervaringen met de methode en

Tabel 1. *Betrouwbaarheid (Cronbachs alfa) en overeenstemming (phi-coëfficiënt) tussen de experts.*

	Betrouwbaarheid		Overeenstemming	
	Ronde 1	Ronde 2	Ronde 1	Ronde 2
COPD/Astma: diagnostiek	0.66	0.76	0.60	0.73
Astma: behandeling	0.71	0.84	0.65	0.81
COPD: behandeling	0.74	0.76	0.71	0.78

opvattingen over de bruikbaarheid van de methode.

Resultaten

De referentiewaarde voor de subtoetsen 'COPD/Astma:diagnostiek', 'Astma: behandeling' en 'COPD: behandeling' is respectievelijk 52.7, 31.9 en 49.6.

Betrouwbaarheid van de beoordelingen en de overeenstemming tussen de experts

In tabel 1 wordt per subtoets de betrouwbaarheid (Cronbachs alfa) en de overeenstemming tussen de experts (phi-coëfficiënt) in de eerste en de tweede ronde weergegeven. De betrouwbaarheid en de mate van overeenstemming nemen voor elke subtoets in de tweede ronde toe. De betrouwbaarheid bereikt de gewenste 0.80 als de subtoetsen in twee ronden door vijf tot acht experts beoordeeld worden (tabel 2).

Tabel 2. *Benodigd aantal experts om betrouwbaarheid (Cronbachs alfa) van 0.80 te bereiken.*

	Ronde 1	Ronde 2
COPD/Astma: diagnostiek	13	8
Astma: behandeling	10	5
COPD: behandeling	9	8

Samenhang geobserveerde en geschatte itemmoeilijkheden

In tabel 3 is te zien dat er een significante ($a_{eenzijdig} < 0.05$) samenhang bestaat tussen de in de tweede expertronde geschatte p-waarden en de geobserveerde p-waarden. Voor de subtoets met de laagste correlatie, 'Astma: behandeling', is bij grafische inspectie te zien dat er sprake is van één 'outlier'. De correlatie tussen de geschatte en de geobserveerde p-waarden bedraagt 0.80 als deze 'outlier' niet wordt meegenomen in de berekening.

Tabel 3. *Samenhang tussen geschatte en geobserveerde p-waarden.*

Subtoets	Pearson correlatiecoëfficiënt
COPD/Astma: diagnostiek	0.835 ($p < 0.01$)
Astma: behandeling	0.601 ($p < 0.05$)
COPD: behandeling	0.827 ($p < 0.01$)

Verschil in slaagpercentages tussen voorbereiders en niet-voorbereiders

In tabel 4 staat per subtoets voor de 'voorbereiders' en de 'niet-voorbereiders' het percentage weergegeven dat beneden de referentiewaarde scoort. Het percentage 'voor bereiders' dat zakt is lager dan het percentage gezakten bij de 'niet-voorbereiders'.

Totale percentage met een score onder en boven de referentiewaarde

In tabel 5 staat het percentage deelnemers, per subtoets, dat onder de absolute en de relatieve referentiewaarde scoort. De absolute referentiewaarde voor de subtoets 'COPD/Astma: diagnostiek' is minder streng

Tabel 4. *Percentage deelnemers dat lager scoort dan de absolute referentiewaarde in de verschillende subgroepen.*

	Voorbereid	Niet voorbereid
COPD/Astma: diagnostiek	3%	27%
Astma: behandeling	10%	70%
COPD: behandeling	16%	53%

Tabel 5. *Percentage deelnemers dat onder de absolute en de relatieve referentiewaarde scoort.*

	Absolute referentiewaarde	Relatieve referentiewaarde
COPD/Astma: diagnostiek	31.5%	47.8%
Astma: behandeling	65.1%	46.7%
COPD: behandeling	48.4%	48.4%

dan de relatieve referentiewaarde. Bij de subtoets 'Astma: behandeling' is de absolute referentiewaarde juist strenger dan de relatieve referentiewaarde. Bij de toets 'COPD: behandeling' komt het percentage dat boven de absolute referentiewaarde scoort overeen met het percentage dat boven de relatieve referentiewaarde scoort.

Praktische uitvoerbaarheid

De benodigde tijd voor de beoordeling – dat wil zeggen het lezen van de instructie, het beoordelen en het invullen van de vragenlijst – loopt sterk uiteen: van 20 minuten tot 95 minuten in de eerste ronde, en van 10 minuten tot 40 minuten in de tweede ronde. Gemiddeld heeft een expert per subtoets 40 minuten nodig voor de eerste en 16 minuten voor de tweede ronde. (Let wel: in de tweede ronde zijn alleen de items beoordeeld waarvan de beoordelingen in de eerste ronde sterk uiteenliepen.) De experts vinden de instructies begrijpelijk. Daarnaast vinden ze de opdracht goed uitvoerbaar. Echter, de meeste experts vinden hun beoordeling arbitrair, wat hen doet twijfelen aan de waarde van de procedure en hun rol als expert.

Beschouwing en conclusie

Onderzocht is de betrouwbaarheid, geloofwaardigheid en de praktische toepasbaarheid van de methode van Angoff voor het bepalen van de referentiewaarde voor de NHG-kennistoets 'COPD/Astma'. Conclusies van het onderzoek zijn dat de experts het verschil in moeilijkheid tussen de items goed kunnen inschatten. De betrouwbaarheid en de mate van overeenstemming tussen de experts na de tweede ronde bereikt vermoedelijk een acceptabel niveau als er twee beoordelaars toegevoegd worden. Gezien het geringe percentage 'voorbereiders' dat onder de referentiewaarde scoort, lijkt de referentiewaarde reëel. De voor de beoordeling benodigde tijd is acceptabel.

Bij dit onderzoek zijn enkele belangrijke kanttekeningen te plaatsen. Ten eerste is de minimaal competente huisarts gedefinieerd als 'normaal goed functionerende huisarts', ofwel 'een huisarts die door nascholing zijn of haar kennis, vaardigheden en praktijk-

voering in voldoende mate op peil houdt'. Deze definitie heeft zowel betrekking op de kennis van het te toetsen gebied als het nascholingsgedrag van de huisarts, terwijl de toets uitsluitend bedoeld is om het kennisniveau van de huisarts te meten. Een argument voor het hanteren van een brede definitie is dat deze brede definitie het voor de experts makkelijker maakt zich een totaalbeeld van de minimaal competente huisarts te vormen. Een tegenargument is dat de gehanteerde definitie geen directe relatie heeft met het domein waarop de toets betrekking heeft. Bovendien is het de vraag of de gebruikte definitie erkend wordt door andere belanghebbenden. Eén expert vond bijvoorbeeld dat de definitie te veel geformuleerd was vanuit de visie van de Afdeling Deskundigheidsbevordering van het NHG. In vervolgonderzoek zal bekeken moeten worden of deze definitie aanpassing behoeft.

Ook de samenstelling van het expertpanel kan verbeterd worden. De experts in het panel zijn voornamelijk expert op inhoudelijk niveau. In de literatuur wordt gesteld dat een goede expert ook voldoende contacten met studenten moet hebben.[5] In vervolgonderzoek zullen meer experts in het panel opgenomen moeten worden, die nauw betrokken zijn bij de nascholing, bijvoorbeeld vertegenwoordigers van de nascholing uit verschillende regio's. Een bijkomend voordeel hiervan is dat door de uitbreiding van het panel de betrouwbaarheid van de referentiewaarde naar alle waarschijnlijkheid zal toenemen.

Een ander discussiepunt is het feit dat de experts aangeven hun beoordeling erg arbitrair te vinden. Echter, het arbitraire karakter is inherent aan de procedure.[11] Belangrijk om hierbij te benadrukken is dat, hoewel het lijkt alsof de beoordelingen erg arbitrair zijn, de gevonden resultaten het tegendeel bewijzen.

Ten slotte is er de vraag waarom er niet gekozen is voor een methode voor de bepaling van de referentiewaarde waarbij gebruik gemaakt wordt van het relevantieoordeel. In bijvoorbeeld de methode van Ebel wordt de experts niet alleen gevraagd de moeilijkheid van de items te schatten maar ook de items te classificeren naar de mate van relevantie.[12] Omdat het ontbreken van bepaalde kennis bij sommige onderwerpen grotere of ingrijpender gevolgen heeft dan bij andere onderwerpen, lijkt relevantie voor de kennistoetsing in de huisartsengeneeskunde van belang. Gebrek aan kennis kan bijvoorbeeld leiden tot foutief handelen in levensbedreigende situaties. Een bezwaar tegen het gebruik van het relevantieoordeel bij de bepaling van de referentiewaarde is dat uit de literatuur blijkt dat er vaak grote meningsverschillen bestaan tussen de experts over de relevantie van een item. De meningen van experts over de moeilijkheid van een item blijken meer overeenstemming te vertonen dan hun meningen over de mate van relevantie ervan. Wellicht kan in een volgend onderzoek gekeken worden of het zinnig is de relevantie bij de bepaling van de referentiewaarde te betrekken.

Samenvattend lijkt de aangepaste methode van Angoff een geschikte methode voor het vaststellen van een absolute referentiewaarde voor screenende kennistoetsing voor huisartsen. De hierboven genoemde discussiepunten dienen dan wel nader onderzocht te worden.

Literatuur

1. Pollemans M. Kennistoetsing bij huisartsen. Maastricht: Universitaire Pers Maastricht; 1994.
2. Dousma T, Horsten A, Brants J. Tentamineren. Groningen: Wolters Noordhoff BV; 1997.
3. Berk RA. A consumer's guide to setting performance standards on criterion-referenced tests. Review of Educational Research 1986;56:137-72.
4. Cizek GJ. Setting passing scores. Educational Measurement: Issues and Practice 1996;15:20-30.
5. Norcini JJ, Shea JA. The credibility and comparability of standards. Applied Measurement in Education 1997;10:39-60.
6. Jansen JJM, Scherpbier AJJA, Kramer AWM, Bloemen JCM, Tan LHC. Toetsing van technische

vaardigheden van huisartsen en huisartsen-in-opleiding: zijn ze geslaagd? In: Spaai GWG, Verweij AMJJ, Remmen R, Dolmans DPHJM, redactie. Gezond Onderwijs-8. Houten/Diegem: Bohn Stafleu Van Loghum; 1999. p. 194-6.
7. Kramer AWM, Jansen JJM, Bloemen JCM, Scherpbier AJJA, Zuithoff P, Tan LHC. Normstelling voor de vaardigheidstoets voor huisartsen-in-opleiding. In: Spaai GWG, Verweij AMJJ, Remmen R, Dolmans DHJM, redactie. Gezond Onderwijs-8. Houten/Diegem: Bohn Stafleu Van Loghum; 1999. p. 210-2.
8. Verhoeven BH, Steeg AFW van der, Scherpbier AJJA, Muijtjens AMM, Verwijnen GM, Vleuten CPM van der. Reliability and credibility of an Angoff standard setting procedure in progress testing using recent graduates as judges. Med Educ 1999;33:832-7.
9. Angoff WH. Scales, norms and equivalent scores. In: Thorndike RL, redactie. Educational measure ment. Washington, DC: American Council on Education; 1971. p. 508-600.
10. Heuvelmans APJM, Sanders PF. Beoordelaars-overeenstemming. In: Eggen TJHM, Sanders PF, redactie. Psychometrie in de praktijk. Arnhem: Cito; 1993. p. 443-69.
11. Glass GV. Standards and criteria. Journal of Educational Measurement 1978;15:237-61.
12. Ebel RL. Essentials of educational measurement. Englewood Cliffs NJ: Prentice-Hall; 1972.

De auteurs

Drs. J. ter Burg, onderwijskundig medewerker, Afdeling Deskundigheidsbevordering, Nederlands Huisartsen Genootschap.
Drs. W.H.M. Emons, assistent in opleiding, Katholieke Universiteit Brabant.
S.S.L. Mol, huisarts-staflid, Afdeling Deskundigheids-bevordering, Nederlands Huisartsen Genootschap.
Dr. L.W.T. Schuwirth, arts, universitair docent, Capaciteitsgroep Onderwijsontwikkeling en Onderwijs-research, Universiteit Maastricht.

Correspondentieadres:
J. ter Burg, afdeling DKB, Nederlands Huisartsen Genootschap, Postbus 3231, 3502 GE Utrecht, tel. 030 288 17 00, fax. 030 287 06 68, e-mail: dkb@nhg-nl.org.

Summary

Introduction: *A performance standard is needed to provide feedback to GPs on their knowledge of the guidelines of the Dutch College of General Practitioners. We examined whether an Angoff procedure is appropriate for test standard setting.*

Method: *Using an Angoff procedure, six experts judged the difficulty of three subtests assessing GPs' knowledge of the diagnosis and management of asthma and COPD.*

Results: *The reliability of the judgments was high. With eight experts it would have been >0.80. The correlation between the p-value estimated by the experts and that observed in 289 GPs who took the knowledge test was between 0.78 and 0.85, which supports the credibility of this method. The feasibility of the procedure (time spent on assessments, clarity of the instructions, et cetera) was high.*

Conclusion: *The Angoff procedure appears to be an adequate method for setting a cut-off score to be used in providing feedback to GPs on their knowledge of the guidelines. (Burg J ter, Emons WHM, Mol SSL, Schuwirth LWT. An Angoff standard setting procedure for screening tests of GPs' knowledge. Dutch Journal of Medical Education 2001;20(1):17-24.)*